Automatic Generation of Training Samples and a Learning Method Based on Advanced MILBoost for Human Detection

Yuji Yamauchi, Hironobu Fujiyoshi
Chubu University
yuu@vision.cs.chubu.ac.jp, hf@cs.chubu.ac.jp

***Abstract*—Statistical learning methods for human detection require large quantities of training samples and thus suffer from high sample collection costs. Their detection performance is also liable to be lower when the training samples are collected in a different environment than the one in which the detection system must operate. In this paper we propose a generative learning method that uses the automatic generation of training samples from 3D models together with an advanced MILBoost learning algorithm. In this study, we use a three-dimensional human model to automatically generate positive samples for learning specialized to specific scenes. Negative training samples are collected by random automatic extraction from video stream, but some of these samples may be collected with incorrect labeling. When a classifier is trained by statistical learning using incorrectly labeled training samples, detection performance is impaired. Therefore, in this study an improved version of MILBoost is used to perform generative learning which is immune to the adverse effects of incorrectly labeled samples among the training samples. In evaluation, we found that a classifier trained using training samples generated from a 3D human model was capable of better detection performance than a classifier trained using training samples extracted by hand. The proposed method can also mitigate the degradation of detection performance when there are image of people mixed in with the negative samples used for learning.**

I. Introduction

Technologies for the automatic detection of humans from images are expected to be implemented in a wide variety of fields such as security and marketing, and studies into improving the precision of human detection have resulted in a large number of proposed methods [1]-[8]. Most of the human detection methods proposed in recent years are based on a consideration of how to capture feature that are suitable for distinguishing humans, such as feature based on the shape of humans [4],[5], feature based on human motion [1]-[3],[6], and feature based on color information [7],[8]. These feature have helped to improve detection performance by picking up on human-like attributes while absorbing factors associated with differences between individuals such as posture, body shape and clothing that can make it difficult to detect humans. However, when the environment in which the training database is collected differs from the scene where the detection system is operated, the human detection performance may be impaired. Solving this problem entails retraining the classifier by collecting data from the environment in which the system operates. However, a great deal of time and effort is needed to prepare separate data sets to train the classifier to detect humans in each scene, making this approach difficult to apply in practice.

As an approach to solving these problems, we have proposed a generative learning method [11] in which modified training samples are generated from a small number of training samples so as to include variations such as changes of scale and the addition of noise that are liable to be measured in real environments, and these samples are used to train the classifier. In reference [9], traffic signs as seen from a vehicle-mounted camera are used to generate samples including effects such as optical blur and motion blur, which are then used for learning. In reference [10], training samples of traffic signs are generated using a generative model that takes account of changes in geometry such as position and rotation, changes in textures such as the background, and changes in color caused by the effects of reflections and the like. However, these techniques are targeted at the recognition of relatively simple two-dimensional patterns. It is difficult to generate training samples by a similar approach for non-rigid objects with complex shapes, including humans.

We therefore propose a generative learning method that uses the automatic generation of training samples from 3D models, together with an advanced MILBoost. In this study, we use a three-dimensional human model to automatically generate positive samples for learning specialized to specific scenes. Negative samples for learning are collected by random automatic extraction from video stream, but some of these samples may be collected with incorrect labeling. When a classifier is trained by statistical learning using incorrectly labeled training samples, this can impair its recognition performance. Therefore, in this study an improved version of MILBoost is used to perform generative learning which is immune to the adverse effects of incorrectly labeled samples mixed in with the training samples.

II. Generation of samples

Fig. 1 shows the procedure of the proposed method as far as training the classifier. In the proposed method, to automatically generate training samples specialized for a specific scene, positive samples are produced by using a 3D human model

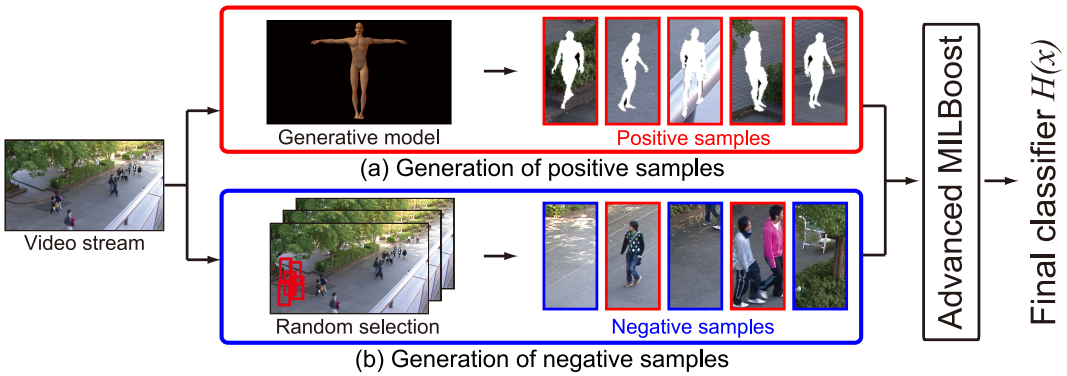

Fig. 1. Generative learning procedure in the proposed method.

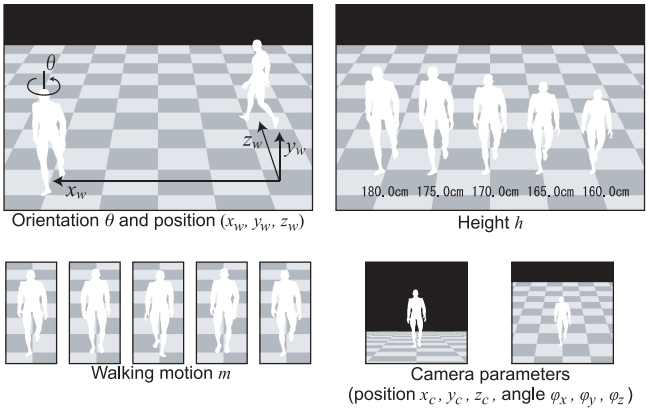

Fig. 2. Adapting the 3D human model to parameters.

to generate human silhouette images (Fig. 1(a)), and negative samples are extracted from the video stream (Fig. 1(b)). These samples are input to an advanced MILBoost algorithm to produce samples that are used to train the classifier.

A. 3D human model

The human model used in the proposed method includes not only a geometrical model but also the hierarchical structure of each part and motion data. The human shape model has 19 parts that are represented in a hierarchical structure. In this study, the parameters of these 19 parts are set up for walking motion so as to reproduce the movements of a walking person. It is also possible to obtain the posture of the human model imaged from any viewpoint as shown in Fig. 2 by applying the following parameters:

- Camera parameters:
  Camera position x_c, y_c, z_c, camera angle ϕ_x, ϕ_y, ϕ_z
- Body shape parameters:
  Height h, orientation θ, position x_h, y_h, z_h
- Texture:
  Background texture T_{bg}, human texture T_{in}

B. Generation of positive samples

To obtain human silhouette images specialized for a particular scene, the parameters of the camera located in the real environment are entered into the 3D human model. In this study, we set up our model with camera parameters that had been obtained beforehand by assuming the camera remains fixed in place. Of the above parameters, the orientation and position of humans are parameters that cannot be determined in advance, so these were made uniformly random. The body height parameter was set to an average value of 171.9 cm based on a statistical survey. The textures of human bodies can be derived by considering details such as clothing, but it is difficult to prepare clothing of a sufficiently broad range of diversity. We considered not applying any textures to the human models, but if learning is performed using untextured samples then the classifier will be trained to expect humans with no internal texture. Therefore in this study we randomly applied textures from natural images that had been prepared beforehand. For the background texture, we used images obtained from the camera mounted. Fig. 3 shows some examples of human silhouette images in a scene where the camera is mounted at a height of $y_c = 6.2\mathrm{m}$ and at a camera angle of $\phi_x = 21°$. Cropped images containing these synthesized human silhouettes in the center were used as positive samples for training.

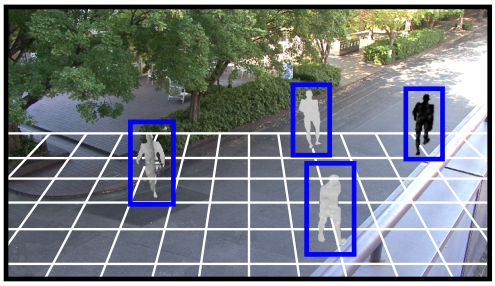

Fig. 3. Examples of silhouette images generated specifically for the real environment(camera position$(x_c, y_c, z_c) = (0\mathrm{m}, 6.2\mathrm{m}, 0\mathrm{m})$, camera angle$(\phi_x, \phi_y, \phi_z) = (21°, 0°, 0°)$).

C. Generation of negative samples

Negative samples were generated by randomly cutting rectangle regions from each frame of video. Although the main purpose is to use frames in which there are no people present, it can be difficult to collect frames consisting entirely of background in places where there is a lot of pedestrian traffic. There is consequently a danger that some of the negative samples collected by cutting out region at random may inadvertently

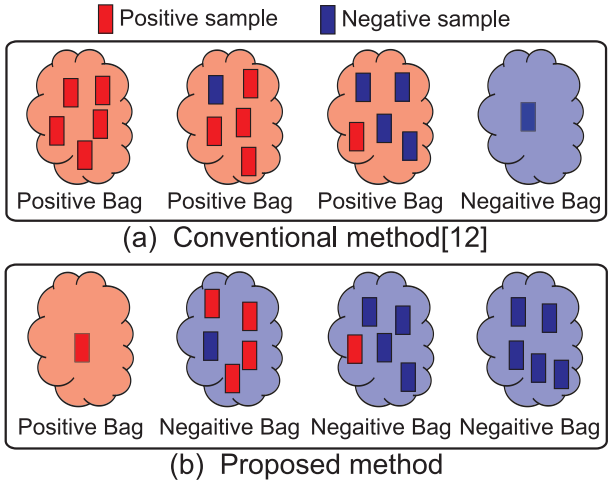

Fig. 4. Composition of bags in the proposed method.

contain images of humans.

In this study, we addressed this problem by training the classifier with an advanced MILBoost algorithm that can cope with the presence of incorrectly labeled samples.

III. Training the Classified with an Advanced MILBoost Algorithm

This section discusses the classifier training method that uses an advanced MILBoost algorithm to solve the problem of incorrectly labeled samples being mixed in with the training samples.

A. MILBoost[12]

Statistical learning methods used for object detection perform learning based on labels applied to training samples. In contrast, Multiple Instance Learning (MIL) works by applying labels to "bags" consisting of multiple samples. In MIL, the classifier is trained based on the labels applied to these bags. Consequently, this approach makes it possible to perform learning based on data that includes unknown samples that have not been labeled. In this study, we used an algorithm called MILBoost [12].

MILBoost is a learning algorithm that introduces the concept of "boosting" into the MIL learning model. Viola et al. have proposed a method for training a face detector efficiently with MILBoost. This method simplifies the collection of positive samples into a positive bag by suitable sampling around the vicinity of faces. Boosting is then used to obtain the class likelihood during sampling for each bag and each sample, and when updating the sample weights, the class likelihoods are used to reduce the weight of incorrectly labeled samples. In this way, the effects of incorrectly labeled samples can be mitigated.

B. An advanced MILBoost algorithm

In the proposed method, the conventional MILBoost algorithm is applied to the problem setting of this study. Here, we first discuss the bag creation method, and then we discuss the

1. Input

Assign a correct label $y \in \{1, 0\}$ to I bags containing J training samples.

2. Initialization

Initialize the weights $w_t(i, j)$ for the training samples.

$$w_1(i,j) = \frac{\text{Bag of class}}{\text{Bag of all}} \tag{1}$$

3. Training

for $t = 1, 2$ to T [T learning iterations] **do**

for $l = 1, 2$ to L [L weak classifier candidates] **do**

Create the probability density function $W_{\pm}$ of weak classifier candidate $h_t(x)$

$$W_+^k = \sum_{i,j:k\in K \wedge y_i=1} w_t(i,j), W_-^k = \sum_{i,j:k\in K \wedge y_i=0} w_t(i,j) \tag{2}$$

Weak classifier $h(x)$

$$h(x) = \frac{1}{2} \ln \frac{W_+^k + \varepsilon}{W_-^k + \varepsilon} (\varepsilon = 1/J) \tag{3}$$

Calculate evaluation value Z_l

$$Z_l = 2 \sum_{k=1}^{K} \sqrt{W_+^k W_-^k} \tag{4}$$

end for

Select the weak classifier candidate $h_t(x)$ with the smallest Z_l

$$h_t(x) = \arg\min_{l\in\mathcal{L}} Z_l \tag{5}$$

Update weights $w_t(i, j)$ of training samples

$$w_{ij} = \begin{cases} -p_{ij} & \text{if } y_i = 1 \\ \frac{p_{ij}\times(-p_i)}{1-p_i} & \text{if } y_i = 0 \end{cases} \tag{6}$$

$$p_i = \prod_{j\in \text{Bag}_i} p_{ij}, \quad p_{ij} = \frac{1}{1+\exp(-H_t(x))} \tag{7}$$

end for

4. Output

Final classifier $H(x)$

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} h_t(x)\right) \tag{8}$$

Fig. 5. Learning algorithm.

advanced MILBoost learning algorithm applied to the problem setting of this study.

1) Bag creation method: In the problem setting of this study, it is not possible to apply the correct label to every single background sample. Consequently, the bag configuration of reference [12] is modified as shown in Fig. 4 . Negative bags are extracted at random from the video. However, there is a possibility that the background of these extracted samples may include images of humans.

2) Learning: The advanced MILBoost learning algorithm is shown in Fig. 5. Apart from the updating of weights for training samples, the advanced MILBoost learning process is shared with real AdaBoost [13], and only differs in terms of the updating of training sample weights.

After selecting a weak classifier, the weight of the training samples is updated so that the incorrectly classified training samples can be correctly classified in the next round. Since MILBoost does not apply class labels directly to the training samples, the weight w_{ij} of the training samples is updated based on the bag label. The weight of samples included in the positive bag is updated by the class likelihood p_{ij} of the samples. A higher class likelihood value indicates things that are more likely to be images of humans, and a lower value

Fig. 6. Examples of training data sets.

TABLE I
EXAMPLES OF TRAINING DATA SETS.

	Positive	Negative
Database 1	INRIA(2,416)	Real environment(12,180)
Database 2	Real environment(2,416)	Real environment(12,180)
Database 3	Generated(2,416)	INRIA(12,180)
Database 4	Generated(2,416)	Real environment(12,180)

indicates things that are more likely to be background images. For samples included in the negative bag, the weights are updated based on the class likelihood p_{ij} of the samples and the class likelihood p_i of the bag. When the sample and bag class likelihoods are low, the sample weight is set to a very large value. When a sample has a low class likelihood and the bag class likelihood is high, the sample weight is set to a large value. Finally, when the sample and bag class likelihood are large, the sample weight is set to a small value. The above process is repeated T times to yield the final classifier $H(x)$.

IV. EXPERIMENTAL EVALUATION

We performed two evaluation experiments to demonstrate the effectiveness of the proposed method. The first experiment demonstrated its effectiveness for generating training samples specialized for a specific scene. In the second experiment, we performed an experimental evaluation to demonstrate the effectiveness of the learning method of the detector with the advanced MILBoost algorithm.

A. A. Experiment 1: Evaluation of automatic generation

1) Experimental overview: Evaluate the effectiveness with the automatic generation of training samples specialized for a specific scene. Comparisons are made between each of the following databases:

- Database 1 : INRIA Pos. + real environment Neg.
- Database 2 : Real environment Pos. + real environment Neg.
- Database 3 : Generated Pos. + INRIA Neg.
- Database 4 : Generated Pos. + real environment Neg.

INRIA Pos. and INRIA Neg. consist of human images and background images included in the INRIA Person Dataset [5]. This database includes diverse background textures and variations of posture, orientation, viewpoint and lighting, making it extremely versatile. Real environment Pos. consists of human images extracted by hand from video captured in the real environment. Generated Pos. consists of human images generated using the 3D human model discussed in Section 2. Real environment Neg. consists of background images extracted manually from video captured in the real environment. The video of the real environment used in this experiment was captured in an outdoor avenue with a large amount of pedestrian traffic. The camera was set at a height of 6.2m with a tilt angle of 21°, and video was captured for approximately one hour. Table I shows the numbers of images in each of the image databases used for learning, and the types of images these database contain. Fig. 6 shows some examples of images in each of the data sets used for learning. For the evaluation database, we used 450 frames selected at random from the video captured in the real environment.

To compare the experimental results, we used a Detection Error Tradeoff (DET) curve. A DET curve is made by plotting False Positives Per Window (FPPW) on the horizontal axis and the miss rate on the vertical axis. Points that are closer to the origin at the bottom left represent a higher detection performance.

2) Experimental results: The DET curves are shown in Fig. 7. First, a comparison of databases 1, 2 and 4 which use the same negative samples shows that the best detection performance is achieved with database 4 which uses samples generated from the human model. This shows that it was possible to generate the appearance of humans to suit the video captured in the real environment. Database 2, which was made using samples extracted manually from video captured in the real environment, achieved results inferior to those obtained by automatic generation. This is thought to be because when human images are extracted by hand, the extraction is performed based on vague criteria which have an adverse effect on the classifier. The lowest detection rates were obtained with database 1, which used a general-purpose database. This is probably because the experimental environment and camera position are different in the INRIA Person Dataset of the training database, so the appearance of people in the sample images was also very different.

Next, a comparison of databases 3 and 4 shows that better results were obtained with database 4, which used background from the video captured in the real environment. This is probably because database 4 uses negative training samples generated from the real environment, thus contributing greatly to the detection performance by producing a classifier specialized for scenes in the real environment.

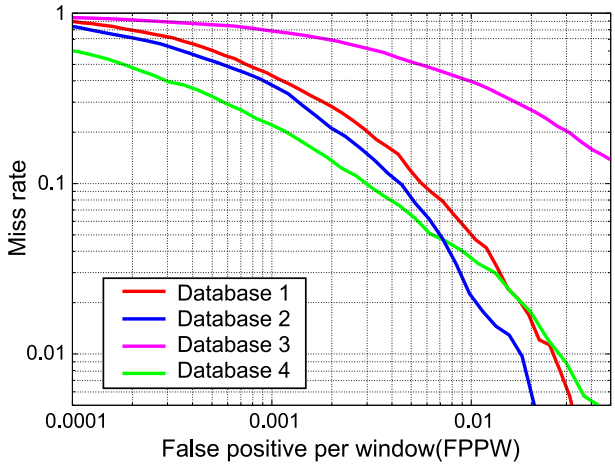

Fig. 7. Experimental results obtained with each training database

B. Experiment 2: Evaluation of the effects of false samples

1) Experimental overview: To evaluate the effectiveness of the proposed method, we compared advanced MILBoost with Real AdaBoost. We trained the classifier with human images deliberately mixed in with the negative samples used for training. For this experiment, we used a database comprising 1,200 positive samples from the INRIA Person Dataset and 4,000 negative samples from the INRIA Person Dataset. To these we added between 0% and 30% of the 1,200 other images in the INRIA Person Dataset that were not used as positive samples. Evaluation was the same as the database used in Experiment 1.

The experimental results were compared in terms of their equal error rate (EER), which is the value at which the miss rate and FPPW become equal.

2) Experimental results: The experimental results are shown in Fig. 8. From these results, it can be seen that as the proportion of human images included in the negative samples increases, the conventional method has a higher EER, while the increase in the EER of the proposed method is suppressed. When the comparison is performed with human images included at a rate of 15%, the EER of the proposed method is 5.8% lower than that of the conventional method. This confirms that the proposed method reduces the adverse effects of human images included in the negative samples when training the classifier.

Fig. 9 shows the output of a strong classifier for background images and human images included in the negative bag, and the changes in the weight of the training samples. From Fig. 9, it can be seen that the human images in the negative bag attain lower weights as the number of learning cycles increases. It can thus be seen that the advanced MILBoost learning algorithm implements learning that is less susceptible to the adverse effects of incorrectly labeled samples.

V. Conclusion

We have proposed a generative learning method that uses the automatic generation of training samples from 3D models, together with an advanced MILBoost. In a specific scene, we were able to train a classified specialized to a real environment by using training samples generated from a 3D human model. Furthermore, by using an advanced MILBoost algorithm, we implemented learning that is less susceptible to the adverse effects of incorrectly labeled samples. In the future, we plan to expand our technique to online learning.

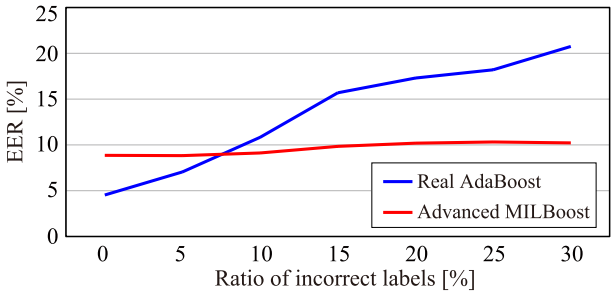

Fig. 8. Change in the ratio of incorrect labels.

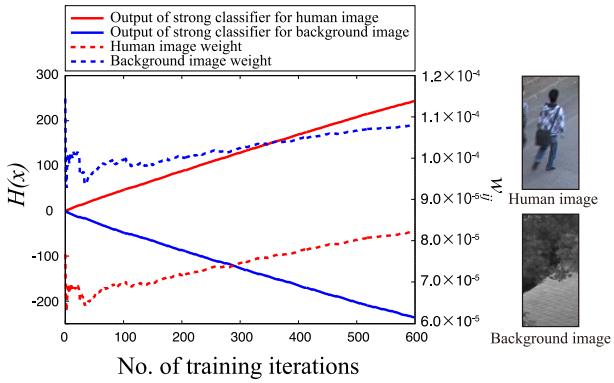

Fig. 9. Changes in classifier output and weight.

References

[1] P. Viola, M. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance", ICCV , Vol. 2, pp.734-741, 2003.

[2] A. Ess and B. Leibe and K. Schindler and L. van Gool. "Moving Obstacle Detection in Highly Dynamic Scenes," ICRA, 2009.

[3] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection", CVPR, 2009.

[4] K. Levi, and Y. Weiss, "Learning object detection from a small number of examples: the importance of good features", CVPR, pp.53-60, 2004.

[5] N. Dalal and B. Triggs: "Histograms of oriented gradients for human detection", CVPR, pp. 886-893, 2005.

[6] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance", ECCV, 2006.

[7] P. Ott, and M. Everingham, "Implicit color segmentation features for pedestrian and object detection", ICCV, 2009.

[8] T. Deselaers, and V. Ferrari, "Global and efficient self similarity for object classification and detection", CVPR, 2010.

[9] M. Noda, T. Takahashi, D. Deguchi, I. Ide, H. Murase, Y. Kojima, T. Naito, "Recognition of Road Markings from In-Vehicle Camera Images by a Generative Learning Method", IAPR Conference on Machine Vision Applications, pp.514-517, 2009.

[10] K. Doman, D. Deguchi, T. Takahashi, Y. Mekada, I. Ide and H. Murase, "Construction of cascaded traffic sign detector using generative learning", International Conference on Innovative Computing Information and Control, pp. 889-892, 2009.

[11] H. Murase, "Generative Learning for Image Recognition ", Information Processing Society of Japan, Vol. 46 No. 15, pp. 35-42, 2005.

[12] P. Viola, John C. Platt and Cha Zhang, "Multiple instance boosting for object detection", NIPS Vol. 18, pp.1419-1426, 2006.

[13] R. E. Schapire, and Y. Singer, "Improved Boosting algorithms using confidence-rated predictions", Machine Learning, pp.297-336, 1999.